Evaluation and Benchmarking Suite for Financial Large Language Models and Agents

Shengyuan Lin[1,2], **Jaisal Patel**[3], **Qinchuan Zhang**[3], **Kaiwen He**[1,3], **Keyi Wang**[1], **Yan Wang**[4], **Matt White**[5], **Kairong Xiao**[6], **Xiao-Yang Liu**[1] *
[1]SecureFinAI Lab, Columbia University, [2]Carnegie Mellon University, [3]Rensselaer Polytechnic Institute, [4]The FinAI, [5]PyTorch Foundation, [6]Business School, Columbia University,
shengyu3@andrew.cmu.edu patelj8@rpi.edu kw2914@columbia.edu wy2266336@gmail.com

Abstract

The financial services industry has witnessed Large Language Models (LLMs) and agents transitioning from the exploration stage to readiness and governance stages. Financial large language models (FinLLMs), such as FinGPT [1] and BloombergGPT [2], have great potential in financial applications, including financial information retrieval, AI tutoring, sentiment analysis, analyzing SEC filings, and agentic trading. However, general-purpose LLMs and agents lack financial expertise and often struggle to handle complex financial reasonings. This paper presents a comprehensive suite for evaluating the FinLLM lifecycle, including the Open FinLLM Leaderboard on HuggingFace. Our collaborative development evolves through three stages: FinLLM Exploration (2023), FinLLM Readiness (2024), and FinAI Governance (2025). The proposed suite serves as an open platform that enables researchers and practitioners to perform both quantitative and qualitative analysis of different FinLLMs and FinAgents, fostering a more robust and reliable FinAI ecosystem.

1 Introduction

Business and finance are high-stakes domains for AI solutions. Evaluating and benchmarking LLMs' gap in financial knowledge and reasoning is critical. It helps pave the path for democratizing financial intelligence to the general public and foster a more robust and reliable FinAI ecosystem.

Several recent works have assessed how LLMs perform on financial certification exams. There are also financial benchmarks, such as FinBen [3] and FinanceBench [4]. However, this interdisciplinary field still lacks a comprehensive framework that systematically evaluates the full lifecycle of FinLLMs and FinAgents, from the early exploration stage to readiness and governance stages. This gap is particularly critical because the financial domain poses unique challenges that demand evaluation methodologies grounded in domain expertise.

To bridge such a gap, our goal is to **explore financial use cases** that reveal both the potential and the limitations of current LLMs across diverse financial scenarios. Currently, the lack of standardized evaluation methods in financial AI creates barriers in which different institutions use varying metrics and evaluation criteria, making it difficult to assess model performance objectively. To address this, we **promote a de facto standard in the financial services industry** through benchmarking and evaluation frameworks.

*Corresponding author.

Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling, 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Figure 1: The evolving FinLLM lifecycle consists of three stages: FinLLM Exploration (2023), FinLLM Readiness (2024), and FinAI Governance (2025), along with a timeline of notable Financial Large Language Models showing the progression from foundational models like BloombergGPT to more recent and multimodal systems.

2 Evolving LLM Lifecycle in Finance: Overview

Over the past three years, SecureFinAI Lab has organized bi-weekly meetings with Linux Foundation and 20+ FinTech companies (industry partners). The topic theme has evolved from LLM Exploration to FinAI Readiness, and to FinAI Governance, as shown in Fig. 1.

- **FinLLM Exploration (2023)**, is the initial exploration into domain-specific financial language models. During this period, researchers and practitioners focused on developing foundational models like BloombergGPT [2] and FinGPT [1], exploring their capabilities in basic financial tasks such as question answering and sentiment analysis.
- **FinLLM Readiness (2024)**, represents a shift toward systematic evaluation and benchmarking. In this stage researchers encouraged the development of comprehensive benchmarks like FinBen [3] and MultiFinBen [5], as well as the emergence of financial agents and specialized APIs. The focus shifted from basic capability demonstration to performance evaluation and real-world applications.
- **FinAI Governance (2025)**, addresses the critical challenges of responsible AI deployment in financial contexts. This stage focuses on identifying and mitigating risks associated with financial LLMs, including hallucination, security vulnerabilities, and regulatory compliance issues. The key initiatives are the AI Governance Framework and the Open Financial LLM Leaderboard for standardized evaluation.

3 Exploration: FinLLM Leaderboard

In 2023, researchers began exploring how large language models could be used in finance. This was the beginning of financial LLMs. People wanted to see if these models could understand financial language and help with financial tasks.

3.1 Timeline for FinLLMs

The evolution of Financial Large Language Models (FinLLMs) can be traced through a timeline of notable milestones. Figure 1 illustrates the rapid shift from foundational financial-domain pretraining to more advanced and multimodal financial AI systems. Beginning with foundational models like BloombergGPT in early 2023, we observe a progression through various specialized financial models, including FinGPT, domain-adapted models, and evolving toward AI governance frameworks.

3.2 Financial Tasks

Financial Tasks with Multimodal Data We compare FinLLMs across multiple task categories including information extraction (IE), textual analysis (TA), question answering (QA), text generation (TG), risk management (RM), forecasting (FO) and Decision-Making (DM). The current 42 financial datasets are organized into seven categories, as given in Table 1. The educational documents of the open FinLLM leaderboard is available on this website.

Table 1: Financial tasks in the open FinLLM leaderboard.

Category	Tasks
Information Extraction (IE)	Named Entity Recognition (NER), Relation Extraction, Causal Classification.
Textual Analysis (TA)	Sentiment Analysis, Hawkish-Dovish Classification, Argument Unit Classification.
Question Answering (QA)	Answering financial questions from datasets like FinQA and TATQA.
Text Generation (TG)	Summarization of financial texts (e.g., reports, filings).
Risk Management (RM)	Credit Scoring, Fraud Detection, evaluating financial risks.
Forecasting (FO)	Stock Movement Prediction based on financial news and social media.
Decision-Making (DM)	Simulating decision-making tasks, e.g., M&A transactions, trading tasks.

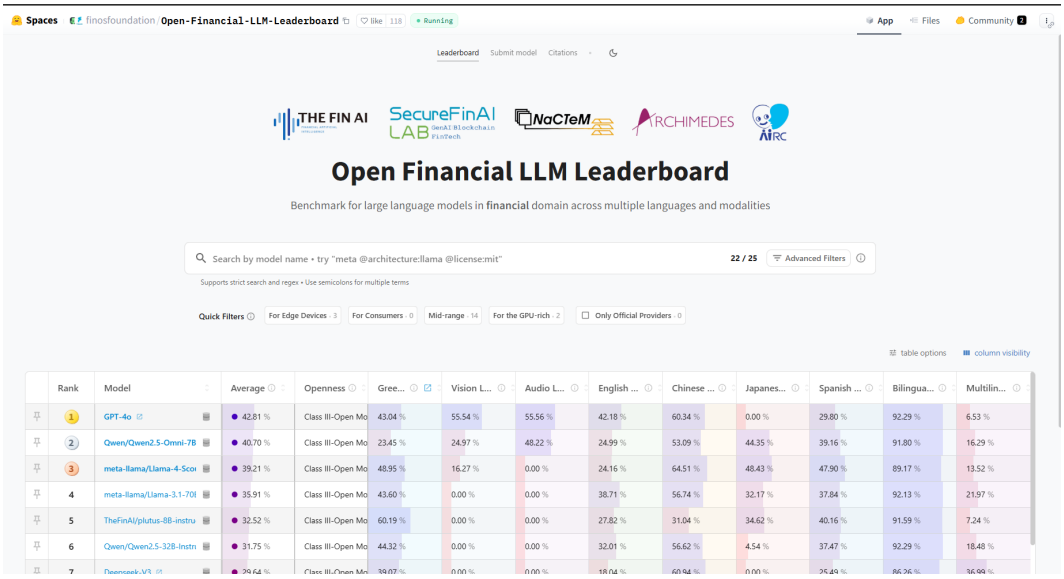

Figure 2: Interface of the Open Financial LLM Leaderboard on Huggingface, showing ranked leaderboard table with multilingual and multimodal financial task scores.

Towards financial readiness Our goal would be to build an open community that pushes financial AI to be ready for real world applications and to build a gateway between academia and industry. The Open FinLLM Leaderboard is similar to established industry standards such as MCP and MOF. We set the benchmark for financial AI readiness, ensuring that innovations in financial language models are both practical and impactful.

3.3 Open FinLLM Leaderboard

The Open FinLLM Leaderboard is an open platform that evaluates and compares FinLLMs and FinAgents across a wide spectrum of financial tasks. It is a collaborative project with the Linux Foundation and Hugging Face. This leaderboard provides a transparent and standardized framework that ranks models based on their (multimodal) performance in areas such as financial reporting, sentiment analysis, and stock prediction. It also serves as an open platform for the community to evaluate, interact with, and compare FinLLMs and FinAgents in real-world scenarios. Beyond numeric scores, we showcase the integration with the FinGPT Search Agent [6], a promising use case of a personalized financial advisor. Users can explore, interact with, and compare models through demos. Additionally, we encourage contributions of models, datasets, and tasks to keep the leaderboard dynamic and responsive to the evolving needs of the financial industry. The leaderboard is continuously evolving, ensuring that it remains up-to-date with the latest FinLLMs and agents and adapts to more professional-level financial tasks. We aim to foster an open collaborative ecosystem for long-term maintenance by following the Model Openness Framework [7] [2] and OpenMDW License[3].

[2] https://isitopen.ai/
[3] https://openmdw.ai/

3.4 Leaderboard Demo

Figure 2 shows the user interface of the Open Financial LLM Leaderboard. The leaderboard on Huggingface is designed to show transparent and comprehensive evaluation results for financial LLMs.

Search Bar and Filtering Tools. Users can search by model name, architecture, and license expression.

Leaderboard Table. The central part is a table containing: rank, model name, average score across all tasks, openness class (Class I, II, III under MOF),

4 Benchmarking Suite: Readiness and Governance

4.1 Financial Agents for Real-World Usages

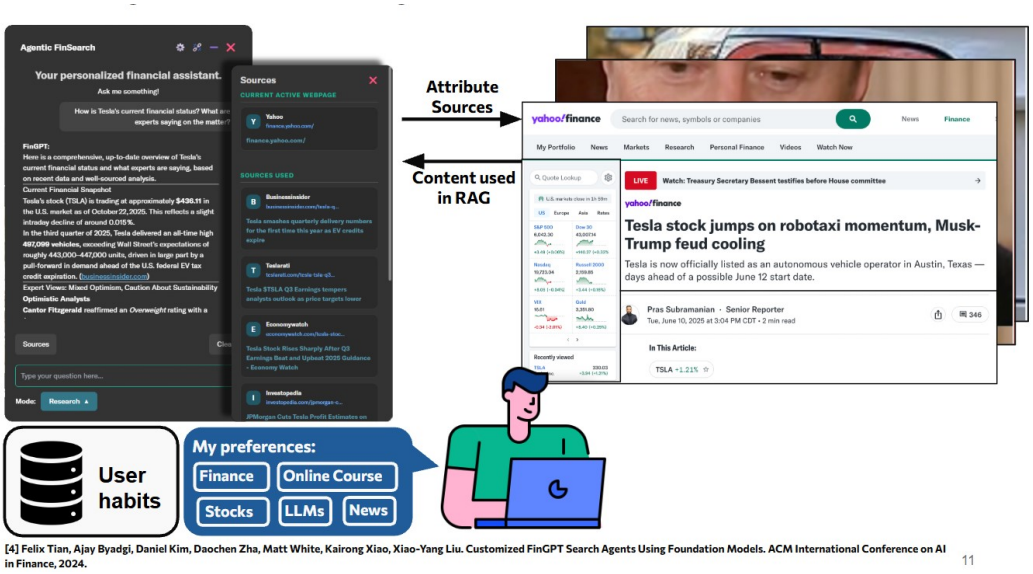

Figure 3: Example 1: Agentic FinSearch for analyzing Tesla stock news from CNBC. The agent scrapes the website, extracts content, and provides key findings with source attribution.

Agentic FinSearch. Agentic FinSearch powered by FinGPT can scrape and analyze sites like YahooFinance and Bloomberg with the integration of RAG [6]. These agents autonomously navigate complex workflows, make informed decisions, and execute multi-step processes. In time-sensitive financial environments, search agents can directly impact profitability with their speed and accuracy.

Unlike traditional financial software that requires manual operation, search agents can autonomously perform tasks such as real-time market data collection, news sentiment analysis, and competitive intelligence gathering.

The Agentic FinSearch is a customizable AI search agent that can scrape and analyze websites, fetch user-specified websites, search local files, and verify sources. It offers two key features: **personalization** and **intelligent search**. The personalization feature learns user habits and preferences from a dynamic database and user feedback, allowing users to maintain a list of preferred websites and select specific LLMs. The intelligent search capability enables the agent to scrape websites or search local files based on user queries, then extract relevant content through RAG and answer queries while providing source attribution.

Fig. 3 demonstrates a practical use case where a user reading a Tesla article on CNBC can directly open the FinGPT window and ask: "based on this article, what are some key findings related to Tesla's stock?" The agent scrapes the user-specified website (CNBC), extracts relevant content, and analyzes it using LLMs to provide a summary of key findings related to Tesla's stock. Users can add frequently accessed websites like CNBC to their preferred websites list for future automated analysis.

Key Features of Agentic FinSearch. Compared to general-purpose AI search solutions, the Agentic FinSearch offers three critical advantages for financial applications:

- **Air-gapped Deployment**. In March 2023, a security incident exposed some ChatGPT users' chat history, raising serious concerns about the leakage risk of users' private financial data. FinGPT can address such vulnerability through air-gapped deployment. It can ensure complete isolation from external networks and prevent data leakage by implementing a local solution for financial institutions to handle sensitive client information and trading strategies.
- **Higher Numerical Accuracy**. Google AI Overview has demonstrated significant accuracy issues in financial contexts: 43% of finance summaries are inaccurate or misleading, and 57% of life insurance information is incorrect. FinGPT can achieve 85% accuracy in financial numerical reasoning tasks, which is higher than Perplexity's 55%. Numerical accuracy is crucial for financial applications in terms of trading decisions and regulatory compliance, where a small mistake can leads to significant losses.
- **Lower Hallucination & Misinformation**. General-purpose AI systems have produced costly errors in financial contexts. Google's AI Overview provided incorrect insurance and Medicare guidance, and Google lost $100 billion in market value at Bard's launch due to a factual error in its demonstration. FinGPT can mitigate risks by fine-tuning with high-quality financial data, retrieval-augmented generation (RAG), and fact-checking responses.

Tutor Agent. The Tutor Agent is an intelligent financial education assistant powered by pretrained large language models that aims to democratize financial education by providing affordable, scalable, and high-quality learning support to the general public. These agents can assist with exam preparation for credit risk management, professional development for learning financial terms, research advising for independent study, and CFA exam preparation with 24/7 access to expert-level explanations and tailored practice questions.

The motivation behind financial AI tutors addresses three key challenges in financial education: **affordability**, **scalability**, and **democratization**. Traditional tutoring services are expensive and inaccessible to many students. AI tutors provide cost-effective deployment using pretrained LLMs, reducing the cost per student through automation. They offer scalable solutions by serving millions of users simultaneously, available 24/7 without geographical limitations, while maintaining consistent quality across all users. Most importantly, they democratize education by breaking down barriers to financial knowledge and enabling self-learning for people from all backgrounds.

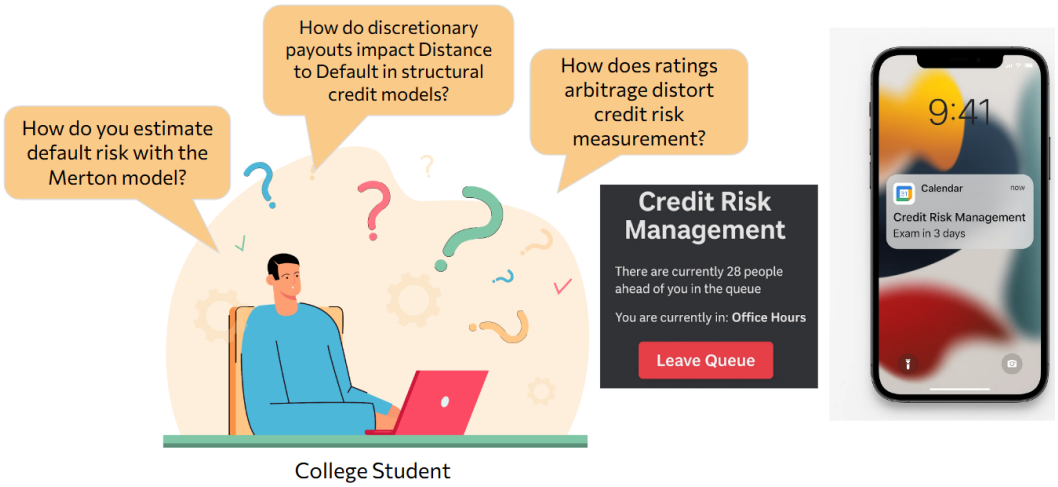

Figure 4: Financial education challenges addressed by AI tutors: affordability, scalability, and democratization of financial knowledge.

Key features include **personalized learning** with 24/7 availability, tailored examples and practice questions, support for multiple educational levels, and step-by-step solution checking. The system

provides **scalable online education** by handling massive student populations, offering real-time answers during lectures, and automated grading using advanced reasoning models.

Practical use cases demonstrate the versatility of financial AI tutors: students preparing for credit risk management exams receive immediate assistance without waiting time; software engineers quickly learn financial terms before meetings with FP&A teams; online students pursuing independent research get guidance on unexplored areas and ongoing feedback; CFA candidates access expert-level explanations and tailored practice questions for complex topics.

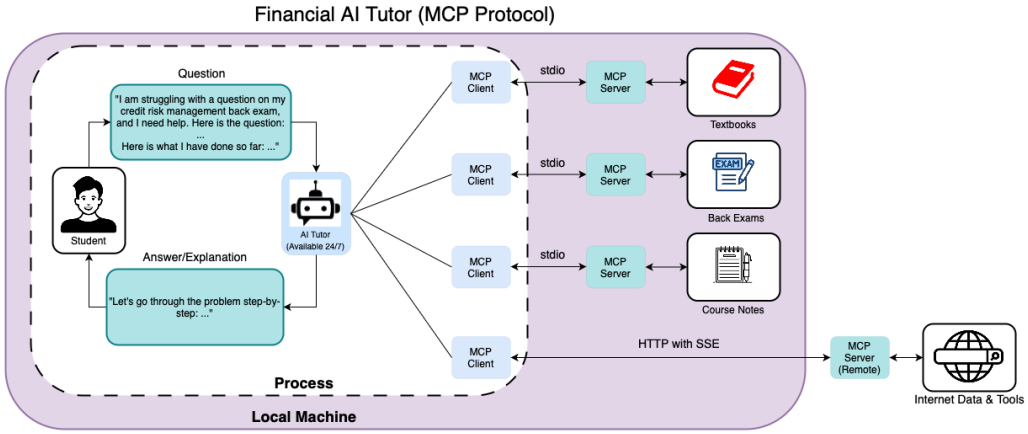

Figure 5: Financial AI tutor framework using MCP Protocol. The tutor provides instant answers to finance questions by accessing trusted resources like textbooks, past exams, and course notes.

Financial AI tutors can access trusted resources like textbooks, past exams, and course notes to provide instant answers to finance questions. They represent a significant advancement in democratizing financial education and making expert-level guidance accessible to a broader audience.

Social Media Monitor: An Example of GameStop Event. The GameStop (GME) short squeeze event in early 2021 highlighted the power of online communities in financial markets. It also generated a massive amount of discussion, questions, and misinformation. Prior work has analyzed the cascading outbreak mechanism of this event using network analysis and LLMs [8]. Financial LLMs can further be used to monitor community discussions, answer complex questions, and provide guidance.

During the GameStop event, retail investors faced unprecedented market conditions and trading restrictions that generated widespread confusion and concern. Figure 6 illustrates how FinLLMs can address questions that emerged from the community during this period.

4.2 FinAI Governance

Generative AI is reshaping financial services by enhancing products, client interactions, and productivity. However, challenges like **hallucinations** and **model unpredictability** make safe deployment complex. The Linux Fundation AI Governance Framework provides a comprehensive collection of risks and mitigations for Generative AI solutions in financial services. These risks can be further identified in the application of FinAgents in financial scenarios, such as private financial data leakage and LLM hallucinations when preparing SEC filings. Retrieval-augmented generation can help reduce hallucinations and ensure financial and regulatory information is up-to-date. Zero-knowledge proof techniques [9] can help protect data privacy and the intellectual properties of LLMs.

5 Conclusion

This paper presents a comprehensive framework for understanding the evolving lifecycle of financial LLMs through three critical stages: Exploration (2023), Readiness (2024), and Governance (2025). We introduce the Open FinLLM Leaderboard as a standardized benchmarking suite that enables systematic evaluation and comparison of financial AI systems. This work contributes to the development of more robust and reliable financial AI systems. Future work will focus on expanding

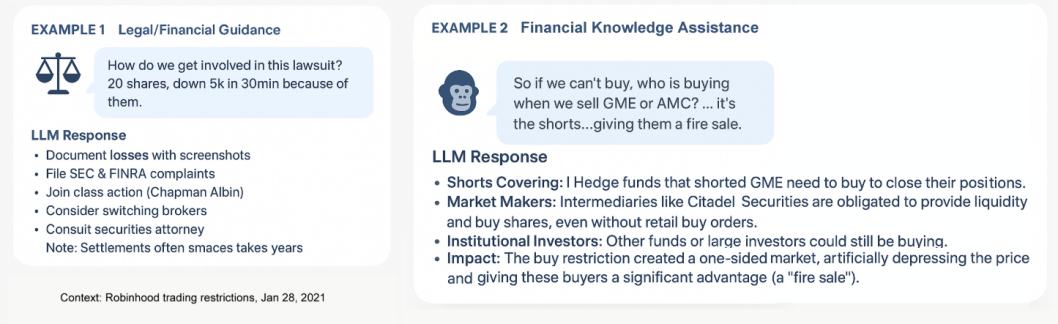

Figure 6: Use cases of FinLLMs: Analyzing community questions from the GameStop squeeze event. The model provides both legal/financial guidance and explains complex market mechanics.

governance frameworks and enhancing evaluation methodologies to address emerging challenges in building a reliable FinAI ecosystem.

References

[1] X.-Y. Liu, G. Wang, H. Yang, and D. Zha, "FinGPT: Democratizing internet-scale data for financial large language models," *Workshop on Instruction Tuning and Instruction Following, NeurIPS*, 2023.

[2] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[3] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng *et al.*, "FinBen: A holistic financial benchmark for large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 95 716–95 743, 2024.

[4] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen, "FinanceBench: A new benchmark for financial question answering," *arXiv preprint arXiv:2311.11944*, 2023.

[5] X. Peng, L. Qian, Y. Wang, R. Xiang, Y. He, Y. Ren, M. Jiang, J. Zhao, H. He, Y. Han *et al.*, "Multifinben: A multilingual, multimodal, and difficulty-aware benchmark for financial llm evaluation," *arXiv preprint arXiv:2506.14028*, 2025.

[6] F. Tian, A. Byadgi, D. S. Kim, D. Zha, M. White, K. Xiao, and X.-Y. Liu, "Customized fingpt search agents using foundation models," in *Proceedings of the 5th ACM International Conference on AI in Finance*, 2024, pp. 469–477.

[7] M. White, I. Haddad, C. Osborne, X.-Y. Y. Liu, A. Abdelmonsef, S. Varghese, and A. L. Hors, "The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence," 2024.

[8] S. Lin, K. Wang, and X.-Y. Liu, "Analyzing cascading outbreak of gamestop event: A practical approach using network analysis and large language models," in *Proceedings of the 5th ACM International Conference on AI in Finance*, 2024, pp. 428–436.

[9] H. Sun, J. Li, and H. Zhang, "zkLLM: Zero knowledge proofs for large language models," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 4405–4419.

[10] F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, and T.-S. Chua, "Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance," *arXiv preprint arXiv:2105.07624*, 2021.

[11] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[12] N. B. Yoash, M. Brief, O. Ovadia, G. Shenderovitz, M. Mishaeli, R. Lemberg, and E. Sheetrit, "Secque: A benchmark for evaluating real-world financial analysis capabilities," *arXiv preprint arXiv:2504.04596*, 2025.

A Financial Agents

Fig. 8 provides examples of how financial agents support analysis and decision-making processes.

Sentiment Analyzer. Bloomberg created BloombergGPT [2] in 2023. It was a 50-billion parameter model trained on financial data. This was one of the first large language models made specifically for finance. It was trained on financial news, reports, and market data. The model learned to understand financial terms and context. The model could analyze **earnings call transcripts**, **financial news articles**, **regulatory filings**, and **financial jargon**.

BloombergGPT demonstrated that financial LLMs can perform sentiment analysis more effectively than general models and proved that domain-specific models are more effective for financial tasks. Financial LLMs could understand the complex language of finance.

Among BloombergGPT's five public financial benchmarks, two tasks (FPB and FiQA SA) target sentiment analysis. In addition, five out of its twelve internal financial benchmarks are sentiment-related, representing nearly half of the total. These statistics illustrate BloombergGPT's prioritization of sentiment analysis and reinforce the central role of sentiment understanding in financial LLMs.

LLMs demonstrate significant potential for generating alpha signals from financial news through sentiment analysis. After retrieving financial information using search agents, users can extract trading signals for informed decision-making. For example, when analyzing CNBC news about Tesla, users can ask: "How is the market reacting to this news?" The LLM can interpret the article, analyze sentiment and tone, assign sentiment scores, and provide rationale. Figure 7 shows how LLMs can conclude that while immediate market reaction appears positive, long-term sentiment remains cautious due to ongoing challenges. This demonstrates LLMs' capability in extracting actionable signals from unstructured financial documents for integration into trading strategies.

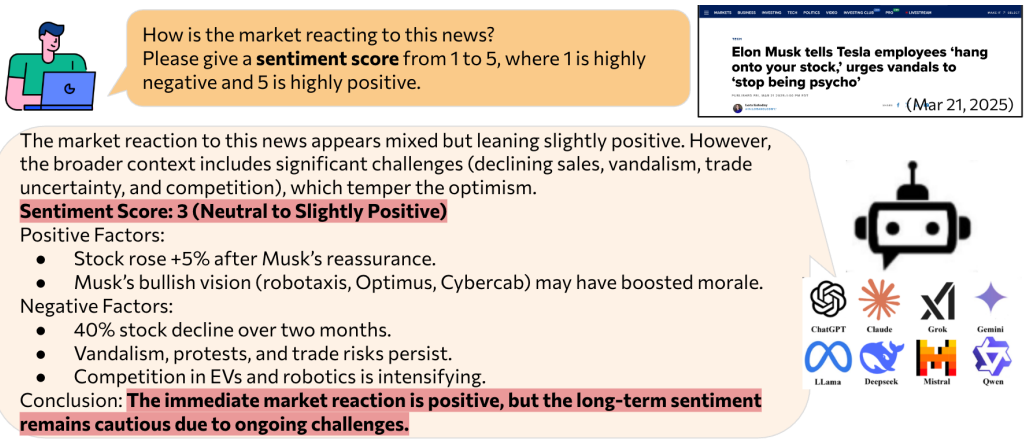

Figure 7: LLM-based sentiment analysis for generating alpha signals from financial news. The system analyzes market sentiment and provides trading signals with rationale.

SEC Analyzer. SEC filings (e.g., 10-K, 10-Q) are the primary way for public companies to disclose information like revenue and cash flows to investors and regulators. SEC filings can be used by FinLLMs for investment and compliance purposes. However, to be effective on these documents, FinLLMs must have strong capabilities in long-context retrieval, numerical reasoning, table extraction and analysis, and temporal understanding of data.

Multiple datasets benchmarking these capabilities have been released. In 2021, TAT-QA [10] introduced 16k+ questions from hybrid tabular and textual data, testing numerical reasoning and table analysis capabilities. After OpenAI released GPT-4 [11], FinanceBench [4] introduced 10k+ QA pairs from SEC filings, which focus on long-context retrieval, numerical reasoning, and table analysis. GPT-4-Turbo showed mixed performance across different retrieval and context settings, motivating research into better long-context and retrieval-augmented approaches. Recently, SECQUE [12] introduced 565 expert-written questions, spanning four key categories: comparison and trend analysis, ratio analysis, risk factors, and analyst insights. SECQUE also contributed SECQUE-judge, an LLM-as-a-judge with strong human alignment for scoring.

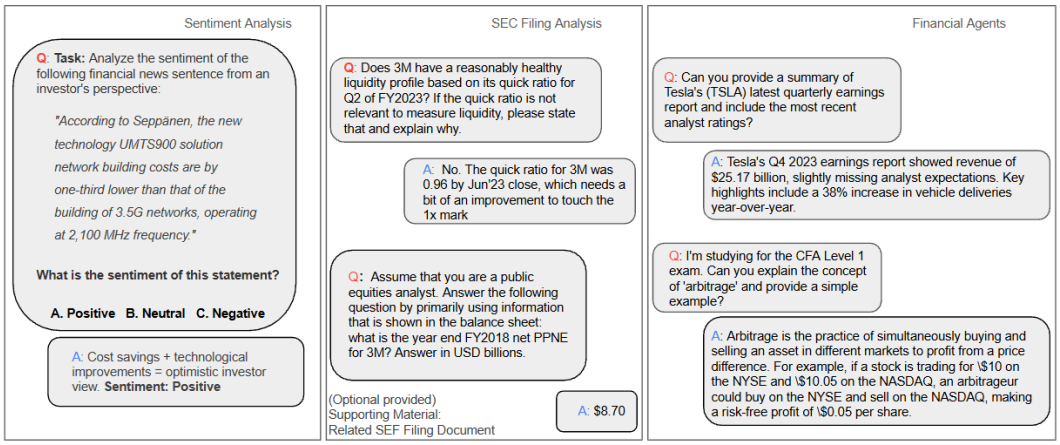

Figure 8: Overview of financial LLM use cases: Sentiment Analysis, SEC Filing Analysis, and Financial Agents, each illustrated with example questions and model answers.

B Model Openness Framework

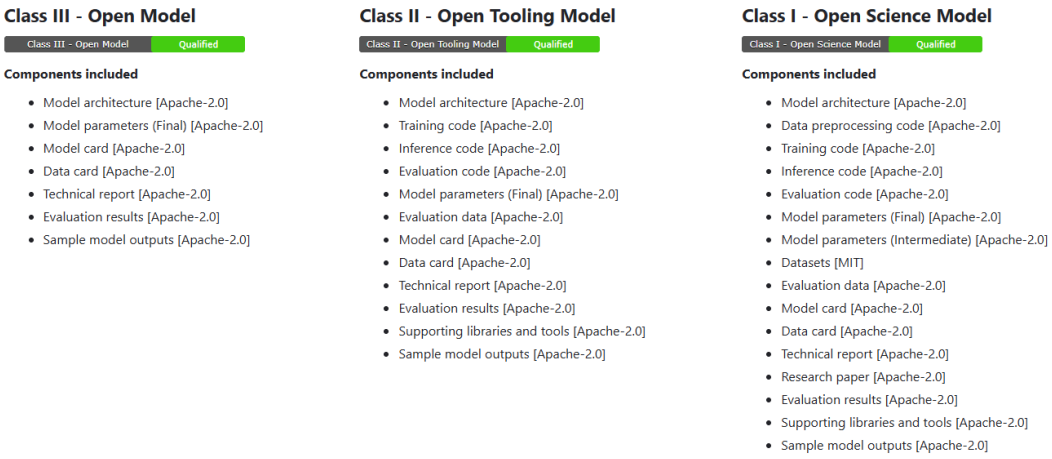

Figure 9: Model Openness Framework (MOF) classification system: Open Model Class 1, Class 2, and Class 3, based on the completeness and accessibility of model components.

The Generative AI Commons at the LF AI & Data Foundation has developed the **Model Openness Framework (MOF)** [7], which evaluates and classifies the completeness and openness of LLMs. With the rise of AI democratization, more and more models are claimed to be open. However, model users often face uncertainty about which specific components are truly open and do not understand the associated licenses. As a result, "openwashing" behavior becomes common among AI models. To address this problem, the MOF identifies 17 components along the lifecycle of model development, including code, data, and documentation, each with suggested open licenses. It classifies the completeness and openness of models into three levels: Class III Open Model, Class II Open Tooling, and Class I Open Science, as shown in Figure 9. With the MOF, model users can better understand what model producers provide, what model components are open, and how to use and distribute models under open licenses. This will enhance the healthy and standardized development of FinAgents built on open models.